ASSIGNMENT 7.1 - DATA MINING

1. Explain the difference between descriptive goals and predictive goals (2%)

Descriptive goal and predictive goal are categorized as discovery goal in data mining whose purpose is to autonomously find patterns in a huge dataset. Descriptive goals need characterization of general properties of the data in the dataset before performing mining. For example, particular descriptor/structure as query (characterization) to search in the data. Predictive goals require prediction to be made using the data in the dataset. Data are extracted from dataset to build model for prediction of unknown compounds.

2. List 2 methods we have covered previously that could be useful in achieving predictive goals (2%)

To achieve predictive goals, we use predictive modeling. The prediction we wish to obtain could be classification nature such as predicting whether the compound is active or inactive, and quantification nature such as predicting the level of activity. Methods in statistics, machine learning, and data mining could be useful. Among them, QSAR (Quantitative Structure-Activity Relationship) model is a concept we use for predicting. Linear regression is one of the methods to predict quantitative problems. It uses the known values in data to build the regression model for prediction of the value of unknown data. But there is a huge assumption in this method: it assumes that there is a linear relationship between the regressor and independent variable. Decision tree learning is also one of the prediction methods. For this ligand-based approach, whether the compounds are active or inactive is known. Then it uses decision tree to train the input data (active/inactive) and build the predictive model.

3. In a heatmap with chemical compounds in rows, protein targets in the columns, and colors in cells indicating activity (red) or inactivity (blue) of compound against targets at the intersections, what descriptors might you use to cluster (i) the compounds and (ii) the targets? (2%)

The choice of descriptor to cluster compounds and targets depends on the goal which we wish to obtain. For example, if the goal is to find the physiology interaction between the compounds and targets, then physicochemical descriptor will be used in clustering (i) and (ii). For the heatmap whose targets are all coming from a certain type of cell line (e.g. cancer cells), then fragmental descriptor will be used to compounds since we want to find out what kind of specific structure are important to the interaction with targets in the cell; the structural descriptor will be used to targets for we want to understand the ligand-based docking situation. If particular structure found in targets has strong interaction with compounds then we get some insight about the feature that causes interaction between them.

4. In the above heatmap, you find a horizontal line of cells colored red. What would this indicate? (2%)

With the compounds in rows and targets in columns, a horizontal line means that this particular compound has activity with all the targets in the assay.

5. Why are pipelining tools useful in data mining? (2%)

In cheminformatics data mining, we not only deal with data analysis, but also concept description (molecular descriptor), data cleaning, integration (cluster or diversity analysis), and selection (for its application in pharmaceutical field). It requires many tools to complete the mining task: visualization, clustering, supervised learning, database querying, statistical analysis, and pattern matching. With the pipeline which could be regarded as graphical scripting methods, these diverse mining tools can be tied together to reach mining purpose.

**ASSIGNMENT 7.2 SCAFFOLD MINNING**

*Data:* Fluorescence Cell-Free Homogeneous Counter Screen to Identify Inhibitors of GFP Chromophore Formation (AID:434968).

*Number of active compounds:* 1764

*Number of inactive compounds:* 349

*Command line in Strip-it:*

* active scaffold mining:

strip-it --input 3237684569133504979.sdf.gz --output active.txt --scaffold scaffold.txt

* inactive scaffold mining:

strip-it --input 176650723242912717.sdf.gz --output inactiv.txt --scaffold scaffold.txt

*Analysis:* using simple python script to calculate the frequency of mined scaffolds in active and inactive compounds. The script is also used to find out the special scaffold which is found in active but not found in inactive.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. Top 10 most frequent scaffold in ACTIVE compound data file:

|  |  |
| --- | --- |
| Scaffold in SMILES | Frequency |
| C1CCC(CC1)C1CCCC1 | 75 |
| C1CCC(CC1)CC1CCCC1 | 62 |
| C1CCCCC1 | 40 |
| C1CCC(CC1)CCC1CCCC2C1CCCC2 | 36 |
| C1CCC2C(C1)CCCC2 | 35 |
| C1CCC(CC1)CC1CCCCC1 | 35 |
| C1CCC(CC1)CC1CCC(C1)C1CCCCC | 33 |
| C1CCC(CC1)C1CCC(C1)C1CCCCC1 | 33 |
| C1CCC2C(C1)CCC2 | 32 |
| C1CCC(CC1)C(C1CCCC1)C1CCCC1 | 31 |

1. Top 10 most frequent scaffold in inactive compounds data file:

|  |  |
| --- | --- |
| Scaffold in SMILES | Frequency |
| C1CCC(CC1)CCC1CCCC1' | 10 |
| C1CCC(CC1)CC1CCCCC1' | 9 |
| C1CCC(CC1)CC1CCCC2C1CCCC2' | 9 |
| C(CC1CCCCC1)CC1CCCCC1' | 7 |
| C1CCC(CC1)C1CCCC1' | 6 |
| C(CCC1CCCC1)CCC1CCC(CC1)CCC1CCCCC1 | 5 |
| C1CCC2C(C1)CCC2 | 5 |
| C1CCC(CC1)CCC1CCCCC1' | 5 |
| C1CCCCC1 | 4 |
| C1CCCC1 | 4 |

1. Top 10 most frequent scaffold found BOTH in active and inactive:

|  |  |
| --- | --- |
| Found BOTH in active and inactive | Frequency |
| C1CCC(CC1)C1CCCC1 | 75 |
| C1CCC(CC1)CC1CCCC1 | 62 |
| C1CCCCC1 | 40 |
| C1CCC(CC1)CC1CCCCC1 | 35 |
| C1CCC2C(C1)CCCC2 | 35 |
| C1CCC(CC1)CC1CCC(C1)C1CCCCC1 | 33 |
| C1CCC(CC1)C1CCC(C1)C1CCCCC1 | 33 |
| C1CCC2C(C1)CCC2 | 32 |
| C1CCC(CC1)C(C1CCCC1)C1CCCC1 | 31 |
| C1CCC(CC1)CCC1CCCC1 | 29 |

The frequency here means how many times the scaffold is found in active and it is also found in inactive. When compare this table to the fist table (scaffold in active), it is shown that almost the scaffold and frequency are the same in the two tables except one scaffold: C1CCC(CC1)CCC1CCCC2C1CCCC2 (frequency in active is 36). As taking the fourth table (scaffold in active but not in inactive, listed below) as reference, it conveys the message that this scaffold, C1CCC(CC1)CCC1CCCC2C1CCCC2, is important in the application of drug discovery since it proves that this scaffold distinguishes the compounds between active and inactive.

1. Top 10 most frequent scaffold found in ACTIVE and NOT found in INACTIVE:

|  |  |
| --- | --- |
| Found in ACTIVE and not found in INACTIVE | Frequency |
| C1CCC(CC1)CCC1CCCC2C1CCCC2 | 36 |
| C1CCC(CC1)CC1CCC(C1)CC1CCCCC1 | 15 |
| C1CCC(CC1)C1CCCC2C1C(CC2)C1CCCCC1 | 13 |
| C1CCC(CC1)C(C1CCC(C1)C1CCCCC1)C1CCC(C1)C1CCCCC1 | 13 |
| C1CCC(CC1)C1CCC2C1C(CCC2)C1CCCC1 | 13 |
| C1CCC(CC1)C1CCCC2C1CCC2C1CCCCC1 | 12 |
| C1CC(CCC2CCCCC2)CC(C1)CCC1CCCCC1 | 12 |
| C1CCC(CC1)CC1CCCC(C1)C1CCCCC1 | 10 |
| C1CCC(CC1)CCC(C1CCCCC1)CC1CCCCC1 | 9 |
| C1CCC(CC1)CC1CC(CCC2CCCC2)C2C(C1)CCCC2 | 9 |

The frequency here means that the specific scaffold is found that many times in active and not found once in inactive. These are the special scaffolds that we may pay attention to for drug discovery because these scaffolds make the compounds to be active rather than inactive with the targets. We can make use of these scaffolds to investigate the feature/critical structure in active compounds.